Dealing With Duplicates

September 29, 2025

[2]:
```
import pandas as pd

data = {
    "Name": ["Alice", "Bob", "Alice", "David"],
    "Age": [25, 30, 25, 40],
    "City": ["NY", "LA", "NY", "Chicago"]
}
```

[3]:
```
df = pd.DataFrame(data)
```

[4]:
```
print("Original DataFrame:")
print(df)
```

```
Original DataFrame:
    Name  Age     City
0  Alice   25       NY
1    Bob   30       LA
2  Alice   25       NY
3  David   40  Chicago
```

[5]:
```
df_cleaned = df.drop_duplicates()
```

[6]:
```
print("\nModified DataFrame (no duplicates)")
print(df_cleaned)
```

```
Modified DataFrame (no duplicates)
    Name  Age     City
0  Alice   25       NY
1    Bob   30       LA
3  David   40  Chicago
```

[7]:
```
df = pd.DataFrame({
    'Name': ['Alice', 'Bob', 'Alice', 'David'],
    'Age': [25, 30, 25, 40],
    'City': ['NY', 'LA', 'SF', 'Chicago']
})
```

[10]:
```
df_cleaned = df.drop_duplicates(subset=["Name"]) # Duplicates are removed only␣
↪based on the Name column while Age and City are ignored for the purpose of␣
↪removing duplicates
```

[9]:
```
print(df_cleaned)
```

```
    Name  Age     City
0  Alice   25       NY
1    Bob   30       LA
3  David   40  Chicago
```

[13]:
```
df = pd.DataFrame({
    'Name': ['Alice', 'Bob', 'Alice', 'David'],
    'Age': [25, 30, 25, 40],
    'City': ['NY', 'LA', 'NY', 'Chicago']
})
df_cleaned = df.drop_duplicates(keep='last') # Keeping the last occurrence of␣
↪duplicate
print(df_cleaned)
```

```
    Name  Age     City
1    Bob   30       LA
2  Alice   25       NY
3  David   40  Chicago
```

[15]:
```
df = pd.DataFrame({
    'Name': ['Alice', 'Bob', 'Alice', 'David'],
    'Age': [25, 30, 25, 40],
    'City': ['NY', 'LA', 'NY', 'Chicago']
})
df_cleaned = df.drop_duplicates(keep=False)  # Dropping all duplicates
print(df_cleaned)
```

```
    Name  Age     City
1    Bob   30       LA
3  David   40  Chicago
```

[16]:
```
df = pd.DataFrame({
    'Name': ['Alice', 'Bob', 'Alice', 'David'],
    'Age': [25, 30, 25, 40],
    'City': ['NY', 'LA', 'NY', 'Chicago']
})
df.drop_duplicates(inplace=True) # Modifying the original DataFrame directly
print(df)
```

```
    Name  Age     City
0  Alice   25       NY
1    Bob   30       LA
3  David   40  Chicago
```

[ ]: